**Fitting Bayesian Ordered Cumulative Logit Hurdle Models**

**Description**

oclhm is used to fit ordered cumulative logit hurdle models using Stan for full Bayesian inference. It can be used to primarily construct a model using data with an ordered outcome that contains informative missingness and include fixed and random effects using lmer syntax in the formula specification. All of distributions supported in Stan are supported in this function with the addition of a “half\_normal” prior option (by default) for the variance component of each parameter. In addition to slope and threshold estimates, the model output includes leave-one-out information criteria, widely accepted information criteria, and intra-class correlations. Additionally, the model output reports model diagnostics via traceplots, mean acceptance rates by chains, and convergence/non-convergence messages for each parameter (based on R-hat).

**Usage**

oclhm(formula, data, chains = 3, cores = 1, iter = 2000, warmup = floor(iter/2), intercept\_prior = "normal(0, 2.5)", slope\_prior = "normal(0, 2.0)", sd\_prior = "half\_normal", fitted = FALSE)

**Arguments**

|  |  |
| --- | --- |
| formula | a two-sided linear formula object describing both the fixed-effects and random-effects part of the model, with the response on the left of a ~ operator and the terms, separated by + operators, on the right. Random-effects terms are distinguished by vertical bars (1|gterm) separating expressions for design matrices from grouping factors. |
| data | an object of class data.frame (or one that can be coerced to that class) containing data of all variables used in the model. |
| chains | number of Markov chains (defaults to 3). |
| cores | number of cores to use when executing the chains in parallel, which defaults to 1 but we recommend using future::availablecores() to determine the number of cores available so one can use as many processors as the hardware and ram allow (up to the number of chains). |
| warmup | a positive integer specifying number of warmup (aka burnin) iterations. This also specifies the number of iterations used for stepsize adaptation, so warmup samples should not be used for inference. the number of warmup should not be larger than iter and the default is iter/2. |
| iter | number of total iterations per chain (including warmup; defaults to 2000). |
| intercept\_prior | prior probability distribution specification for intercepts/thresholds using stan syntax. Defaults to normal(0, 2.5). |
| slope\_prior | prior probability distribution specification for slopes using stan syntax. Defaults to normal(0, 2.0). |
| sd\_prior | prior probability distribution specification for variance component of parameters above using stan syntax. Defaults to half\_normal (see details below). |
| fitted | Logical. If TRUE the 95% credible interval and posterior median of the fitted logit values are reported. |

**Details**

The oclhm function does all the work of fitting an ordered cumulative logit hurdle models using Stan for full Bayesian inference. The steps are roughly as follows:

1. Translate the Stan model to C++ code.
2. Compile the C++ code into a binary shared object, which is loaded into the current R session (an object of S4 class stanmodel is created).
3. Draw samples and wrap them in an object of S4 class stanfit.
4. Summarize samples, report diagnostics, and wrap them in an object of list class.

The function returns a traditional Bayesian summary of the desired estimates, as well as some additional measures. Additionally, the original Stan model (Stan\_model) is saved to the global environment (as Stan\_model) and can be used with methods such as print, summary, and plot to inspect and retrieve the results of the fitted model.

**General formula structure**

The formula argument accepts formulas of the following syntax:

response ~ pterms + (1|gterms)

The pterms part contains effects that are assumed to be the same across observations. We call them 'population-level' or 'overall' effects, or (adopting frequentist vocabulary) 'fixed' effects. The optional gterms part may contain effects that are assumed to vary across grouping variables specified in group. We call them 'group-level' or 'varying' effects, or (adopting frequentist vocabulary) 'random' effects.

In addition to main effects, fixed and random interaction effects can be fit using “:” notation only (e.g., fixed - pterm1:pterm2 *not* pterm1\*pterm2; random – (1|gterm1:gterm2) *not* (1|gterm1\*gterm2)).

**Data**

Data should be in long format (as it would be for use in lm or anova). The response must be ordered and contain informative missingness to be dealt with the hurdle component of the model. Ideally, the user has the ordered outcome specified numerical as “1, 2, 3…” and the informatively missing data as “0”. Character vectors are currently not supported. Predictor variables (fixed and random) may be continuous, discrete, or character vectors (which will be treated as factors).

**Priors**

Priors must be specified as a string using Stan notation. A complete overview on possible prior distributions is given in the Stan Reference Manual available at <https://mc-stan.org/>.

The default intercept\_prior and slope\_prior distributions provided are somewhat conventional and weakly informative but may be changed to suit the researcher’s preference. For the sd\_prior, the use of half\_normal(0, 2.5) for the default variance component prior allows to the user to specify a truncated normal distribution since this option is not explicitly available in Stan (in Stan syntax: normal\_lpdf(vc | 0, 2.5) - 1\*normal\_lccdf(0 | 0, 2.5)).

**Value**

Three objects are saved after using oclhm: the main model summary and diagnostics (name: user specified, structure: list), the original Stan model (name: Stan\_Model, structure: stanfit), and the model formula (name: formula\_Model, structure: list).

The main model summary object is a list and contains the following components:

|  |  |
| --- | --- |
| Summary | posterior mean, standard error, standard deviation, 95% credible interval, effective sample size, and Rhat for each fixed effect as well as threshold estimates. |
| WAIC | widely accepted information criteria for model comparison. |
| LOOIC | leave-one-out information criteria for model comparison. |
| ICC | 95% credible interval of intra-class correlation for each variance component. |
| Mean Accept. Stat by Chain | mean acceptance rate for each chain. |
| Convergence | convergence messages for each parameter in the order specified. |
| Traceplot | chain mixing diagnostic plots for each parameter. |
| Fitted | the 95% credible interval and posterior median of the fitted conditional and unconditional logit values are reported. |

**Author(s)**

Carson Keeter [keeterc1@gmail.com](mailto:keeterc1@gmail.com)

Pavel Chernyavskiy [pchern@virginia.edu](mailto:pchern@virginia.edu)

Traci Kutaka [Traci.Kutaka@du.edu](mailto:Traci.Kutaka@du.edu)